scientific data

Check for updates

OPEN

DATA DESCRIPTOR

# A high-quality chromosome-level genome assembly and annotation of the giant freshwater prawn (*Macrobrachium rosenbergii*)

Lingyun Yu[1,4], Haiyang Liu[1,4], Jie Wei[1], Wanchang Zhang[2], Kunhao Hong[1,3], Huarong Li[1,3], Qiaoyan Zhou[1], Qiyao Su[1,3], Bai Liufu[1,3], Xinping Zhu[1]✉ & Yakun Wang[1]✉

**The giant freshwater prawn, *Macrobrachium rosenbergii*, is native to Southeast Asia and is used in aquacultural practices worldwide. It is considered advantageous because of its rapid growth, high nutritional value, and economic benefits. As one of the three major freshwater aquaculture shrimp sources in China, a high-quality genome resource is of great significance for promoting the germplasm improvement of varieties. This study presents a high-quality chromosome-level genome assembly of *M. rosenbergii* that was generated by combining PacBio, MGI, and Hi-C reads. The assembled genome was 2.96 Gb in size, with a contig N50 of 0.64 Mb and a scaffold N50 of 55.76 Mb, which was positioned on 59 pseudo-chromosomes. The Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis for genome assembly reached 94.37%. In total, 27,111 protein-coding genes were identified, of which 25,470 were functionally annotated. These results provide a foundation for future research into adaptive evolution, genomics, and molecular breeding in *M. rosenbergii*.**

## Background & Summary

The giant freshwater prawn *Macrobrachium rosenbergii* is one of the world's most important commercial freshwater prawn species because of its tender meat, fresh taste and nutrients[1]. The scale of *M. rosenbergii* aquaculture has continued to expand to meet increasing consumer demand. However, challenges such as the frequency of disease and germplasm degradation limit the further development of this industry[2–4]. Thus, there is a need to improve the quality of the available germplasms and selective breeding methods to ensure the sustainable development of the *M. rosenbergii* aquaculture industry[5].

Genomic data can provide a comprehensive understanding of germplasm traits. In recent years, efforts to improve germplasm in aquatic animals have advanced greatly due to the development of methods that can decipher genomic information, such as the development and application of markers associated with sex, growth, and other traits based on genomic association analysis and the use of gene modification technologies to develop new economically valuable varieties[6–10]. Aquatic animals are crucial for aquaculture production, and the genomic information for crustaceans such as *Penaeus vannamei*[11], *Penaeus monodon*[12], *Procambarus clarkii*[13], *Macrobrachium nipponense*[14], *Cherax quadricarinatus*[15], *Antarctic krill*[16], *Portunus trituberculatus*[17], and *Eriocheir sinensis*[18] has been resolved. The resulting data have facilitated the development of respective germplasm resources.

Although transcriptomic analysis and other modern methods have been used to explore the genetic resources related to sex determination and differentiation, disease resistance, salinity tolerance[19–21], the adult male genome information has been published[22], however, the overall progress of germplasm improvement in *M. rosenbergii* has remained relatively slow because of the lack of high-quality female genomic resource[5]. To address this issue, we generated a chromosome-level genome assembly of *M. rosenbergii* with a contig N50 of 0.64 Mb using PacBio

[1]Key Laboratory of Tropical & Subtropical Fishery Resource Application & Cultivation of Ministry of Agriculture and Rural Affairs, Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, 510380, China. [2]College of Life Science, Nanchang University, Nanchang, 330031, China. [3]College of Fisheries and Life Science, Shanghai Ocean University, Shanghai, 201306, China. [4]These authors contributed equally: Lingyun Yu, Haiyang Liu. ✉e-mail: zhuxinping_1964@163.com; wykzkyky@163.com

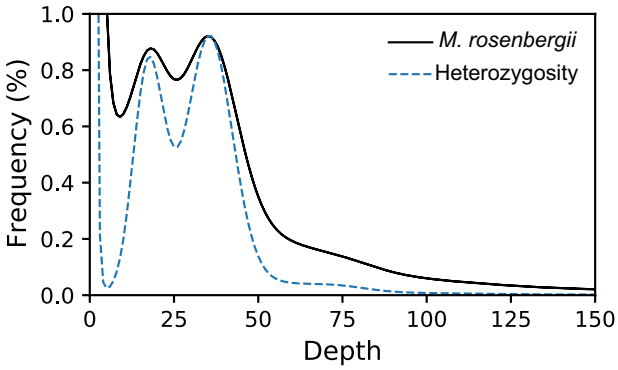


**Fig. 1** Frequency distribution of the 17-mer analysis of the *Macrobrachium rosenbergii* genome. The x-axis represents the K-mer depth and the y-axis represents the frequency of the K-mer corresponding to the depth.

High-Fidelity (HiFi) and chromosome conformation capture (Hi-C) sequencing technologies. Consistency of the conserved core genes (BUSCO scores) and the genome confirmed the continuity and accuracy of the assembly, which represents a high-quality genome assembly to date for *M. rosenbergii*. The resulting high-quality genomic data provide an excellent genetic resource for future efforts to improve *M. rosenbergii* germplasm.

## Methods

**Sampling and sequencing.** A sexually mature female *M. rosenbergii* individual was obtained from a breeding facility at the Pearl River Fisheries Research Institute of the Chinese Academy of Fisheries (Guangzhou, China). The individual was anaesthetized and the heart, hepatopancreas, gill, eye stalks, brain, ovaries, muscles, stomach, and intestine were obtained. The hemolymph was also extracted using a 1-mL syringe. The tissues were rapidly preserved in liquid nitrogen and then stored at −80 °C until further analysis. Muscle tissue was used for high-molecular-weight genomic DNA (gDNA) extraction using a microtissue kit (OMEGA). DNA quality was assessed using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA).

After interruption with g-TUBEs (Covaris, USA), the target DNA fragments were purified, subjected to damage and end repair, and stem-loop sequencing junctions were then attached at both ends. A BluePippin cutter (Sage Science, USA) was used to screen target fragments, and an Agilent 2100 Bioanalyzer (Agilent technologies, USA) was used to detect the size of library fragments. High-quality library for MGI sequencing was constructed by MGIEasy DNA Library Prep Kit V3.1 (MGI Tech, China). DNA sequencing, genomic survey and Hi-C were performed on the MGISEQ T7 platform (MGI, China). PacBio SMRT sequencing[23], and Hi-C libraries were generated using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, USA) on the Pacific Biosciences Sequel II platform in CCS mode and the GrandOmics Hi-C kit (DpnII restriction enzyme; GrandOmics, China), respectively, according to the manufacturer's instructions. Total RNA extracted from 10 tissues obtained above using a TRIzol kit (Invitrogen, Shanghai, China) were mixed and then subjected to RNA-seq on the MGISEQ T7 platform (MGI, China).

**Quality control of raw sequencing data.** The original image data of MGI sequencing were converted into raw sequence data via base calling. All the raw data were then filtered using fastp (v0.23.2)[24] (parameters: -n 0 -f 5 -F 5 -t 5 -T 5 -q 20), and the following filtering criteria were applied: read splice sequences were removed, inaccurate bases at both ends of the reads were truncated, five bases at the left and right ends, respectively, were intercepted, and the reads containing N were removed. When more than 20% of the base mass fraction of a read was $< 20$, the pair of reads corresponding to that read were discarded. FastQC[25] (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.; parameter: -extract) software was used as a quality control system for the clean data. If quality control was achieved, subsequent analyses were performed. Quality control for raw data of PacBio SMRT sequencing was performed using SMRT Link V12.0[26] (default parameters) to filter out fail reads

One hundred thousand reads after QC were randomly compared to the nt data using blastn. The read distribution in the nt library and species distribution were counted to assess genomic data contamination.

**Genome size survey and heterozygosity estimation.** To understand the genomic characteristics of *M. rosenbergii*, K-mer analysis was performed using MGISEQ T7 DNA data with 202.7 Gb of clean reads prior to genome assembly to estimate the genome size and heterozygosity. Briefly, the 17-mer sequence files were counted using Kmc[27]. Skew normal distribution and negative binomial models were used to fit the K-mer data, then, the genome size and heterozygosity were evaluated using the FindGSE[28]. The final genome size of *M. rosenbergii* was approximately 3.03 Gb, and the heterozygosity and repetitive sequence content were approximately 1.40% and 32.6%, respectively (Fig. 1). The visualization result for heterozygosity and repetitive sequence was presented in Python 3 software[29] with the matplotlib package.

**Chromosome level genome assembly and evaluation.** Hifiasm v0.19.8[30] was used to assemble the sequencing data with 152.2 Gb of PacBio HiFi reads into contigs. Then, after removal of redundant contigs based on read depth distribution and sequence similarity by Purge_dups[31], the 333.14 Gb of Hi-C sequencing data was aligned to contigs using bowtie2[32]. Then, contig clustering was performed using the LACHESIS software[33]. In

| Chromosome | Length(bp) | Number of Contigs | Chromosome | Length(bp) | Number of Contigs |
|---|---|---|---|---|---|
| Chr01 | 110,474,880 | 284 | Chr31 | 44,475,496 | 187 |
| Chr02 | 90,304,994 | 144 | Chr32 | 43,940,665 | 207 |
| Chr03 | 89,487,134 | 196 | Chr33 | 43,723,883 | 175 |
| Chr04 | 89,302,480 | 199 | Chr34 | 43,181,680 | 89 |
| Chr05 | 87,816,847 | 183 | Chr35 | 41,456,909 | 116 |
| Chr06 | 86,996,058 | 579 | Chr36 | 40,747,293 | 143 |
| Chr07 | 85,005,314 | 165 | Chr37 | 40,594,584 | 62 |
| Chr08 | 82,891,011 | 189 | Chr38 | 37,808,965 | 127 |
| Chr09 | 78,829,559 | 182 | Chr39 | 36,546,148 | 118 |
| Chr10 | 73,213,210 | 193 | Chr40 | 36,108,234 | 67 |
| Chr11 | 71,621,420 | 156 | Chr41 | 35,960,449 | 77 |
| Chr12 | 71,515,171 | 196 | Chr42 | 35,490,450 | 115 |
| Chr13 | 70,152,309 | 183 | Chr43 | 35,293,769 | 72 |
| Chr14 | 69,880,110 | 151 | Chr44 | 33,539,101 | 69 |
| Chr15 | 66,802,178 | 240 | Chr45 | 31,679,836 | 106 |
| Chr16 | 61,117,927 | 207 | Chr46 | 30,401,081 | 92 |
| Chr17 | 60,416,233 | 157 | Chr47 | 28,172,603 | 148 |
| Chr18 | 57,835,642 | 197 | Chr48 | 28,031,469 | 63 |
| Chr19 | 56,089,517 | 91 | Chr49 | 27,698,305 | 78 |
| Chr20 | 55,742,444 | 174 | Chr50 | 27,606,729 | 107 |
| Chr21 | 55,450,205 | 122 | Chr51 | 24,723,796 | 87 |
| Chr22 | 54,261,800 | 195 | Chr52 | 24,343,172 | 124 |
| Chr23 | 53,277,743 | 164 | Chr53 | 16,381,747 | 42 |
| Chr24 | 50,637,762 | 152 | Chr54 | 15,976,100 | 30 |
| Chr25 | 50,632,732 | 158 | Chr55 | 14,156,513 | 108 |
| Chr26 | 49,681,548 | 148 | Chr56 | 12,310,521 | 85 |
| Chr27 | 49,362,752 | 141 | Chr57 | 7,137,405 | 33 |
| Chr28 | 48,907,357 | 55 | Chr58 | 5,221,204 | 52 |
| Chr29 | 48,123,933 | 280 | Chr59 | 5,114,107 | 31 |
| Chr30 | 47,004,031 | 109 | Total | 2,870,656,515 | 8,400 |

**Table 1.** Statistics of 59 chromosomes of *M. rosenbergii* genome.

total, 3012.15 Mb of the sequence was localized to 59 chromosomes, which accounted for 99.74% of the total length. Contig order and orientation were performed using interactions, and contig sequences with determined order and orientation were concatenated by adding 100 Ns to obtain the final chromosome-level genome sequences, which yielded a chromosome mount rate of 96.85% (Table 1, Fig. 2). The completeness of genome assembly and uniformity of sequencing coverage were assessed using the BUSCO[34] single-copy gene database and minimap2 software[35] with the default parameters and compared with the assembled genome. Finally, we obtained a chromosome-level genome assembly of 2.96 Gb with a contig N50 of 0.64 Mb and a scaffold N50 of 55.76 Mb (Table 2). The genome assembly contained 59 chromosomes ranging in length from 5.11 to 110.47 Mb, with an average length of 47.51 Mb (Table 1). The assembled genome score was 94.37% by BUSCO analysis (Table 3).

**Repetitive sequence annotation.** Repeat elements of *M. rosenbergii* were comprehensively identified using Ab Initio and homology-based strategies with GMATA v2.2[36] and Tandem Repeat Finder (TRF, v 4.07)[37]. GMATA identified the simple sequence repeats (SSRs), and TRF recognized all tandem repeat elements. TEclass[38] was used to classify the unknown repeated sequences, and RepeatMasker[39] was used to search for known and novel transposable elements (TEs) by mapping sequences against the de novo repeat library and Repbase TE library. The overlapping transposable elements belonging to the same repeat class were collated and combined. Overall, 53.16% of the assembled *M. rosenbergii* genome contained repetitive sequences (Fig. 3). Among the TEs, DNA elements (DNAs), long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs) covered 15.86%, 13.21%, 0.69%, and 14.48% of the genome, respectively (Table 4).

**Protein-coding genes prediction.** STAR software[40] was used to compare the clean data with the reference genome after quality control of the raw transcriptomic data which was filtered by using fastp(v0.23.2)[24]. The StringTie was used to assemble the transcript sequencing data into de-novo transcript models. The Program to Assemble Spliced Alignments (PASA) software[41] was then used to make gene predictions for the obtained sequences.

Based on the protein sequence information of neighboring species, the corresponding protein information was compared with the genome using GeMoMa[42], and the prediction results for all homologous species were

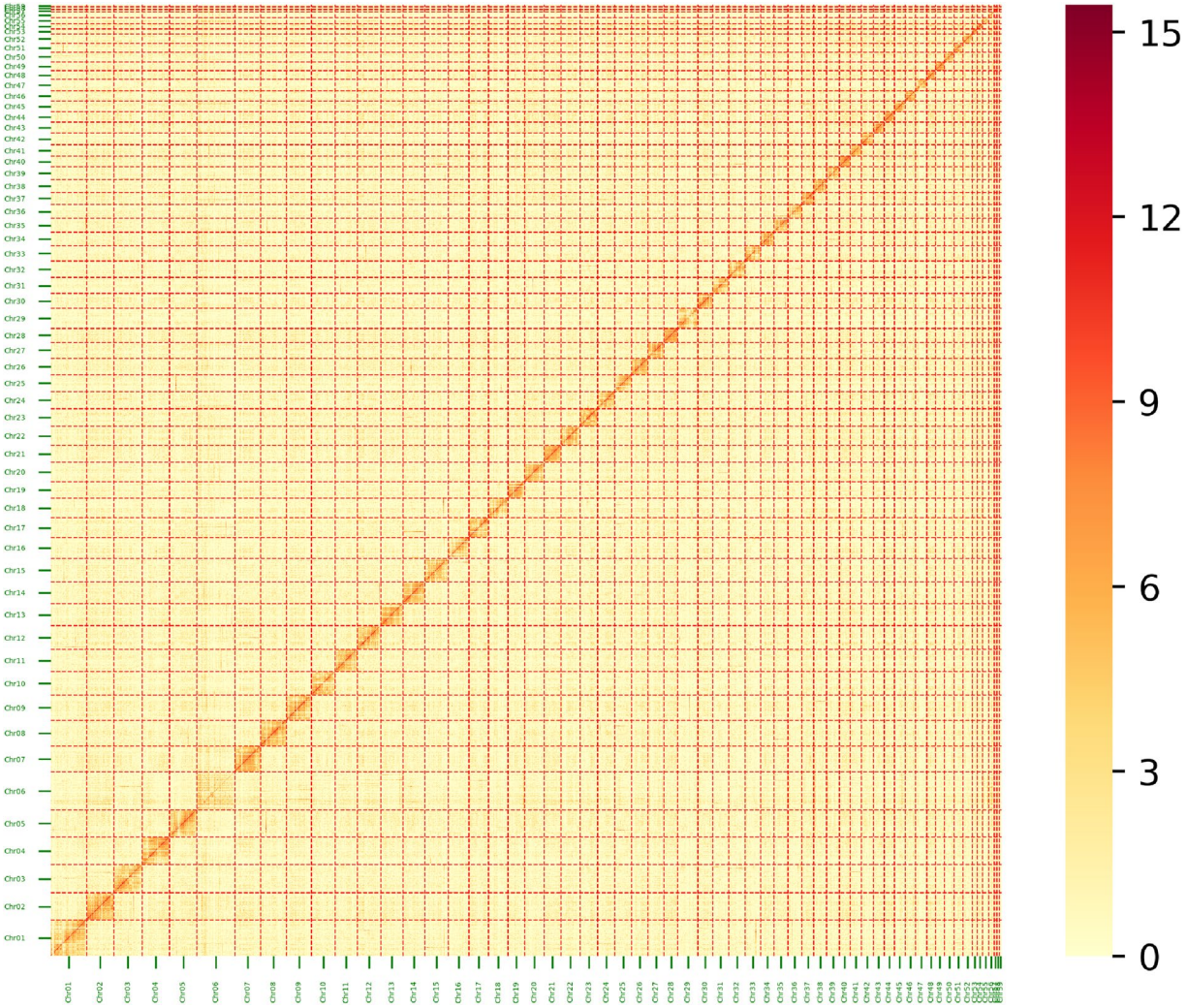


**Fig. 2** Hi-C interaction heatmap of *Macrobrachium rosenbergii*. The map shows the scaffolded and independently assembled chromosomes at high resolution.

| Contig Assembly | | | | Hi-C assembly | | | |
|---|---|---|---|---|---|---|---|
| **Stat Type** | **Contig Length (bp)** | **Stat Type** | **Contig Number** | **Stat Type** | **Scaffold Length (bp)** | **Stat Type** | **Scaffold Number** |
| N50 | 641,689 | L50 | 1,055 | N50 | 55,759,744 | L50 | 20 |
| N60 | 433,711 | L60 | 1,638 | N60 | 50,648,432 | L60 | 25 |
| N70 | 286,364 | L70 | 2,504 | N70 | 43,961,265 | L70 | 32 |
| N80 | 188,181 | L80 | 3,822 | N80 | 36,557,848 | L80 | 39 |
| N90 | 122,032 | L90 | 5,832 | N90 | 28,037,669 | L90 | 48 |
| Total | 3,043,343,653 | Total | 9,697 | Total | 2,964,106,244 | Total | 948 |

**Table 2.** Characteristic of the assembled *M. rosenbergii* genome.

| Type | Number | Percent(%) |
|---|---|---|
| Complete BUSCOs (C) | 957 | 94.37 |
| Complete and single-copy BUSCOs (S) | 951 | 93.88 |
| Complete and duplicated BUSCOs (D) | 6 | 0.59 |
| Fragmented BUSCOs (F) | 12 | 1.18 |
| Missing BUSCOs (M) | 44 | 4.34 |
| Arthropoda | 1,013 | 100 |

**Table 3.** BUSCO statistics for the *M. rosenbergii* genome.

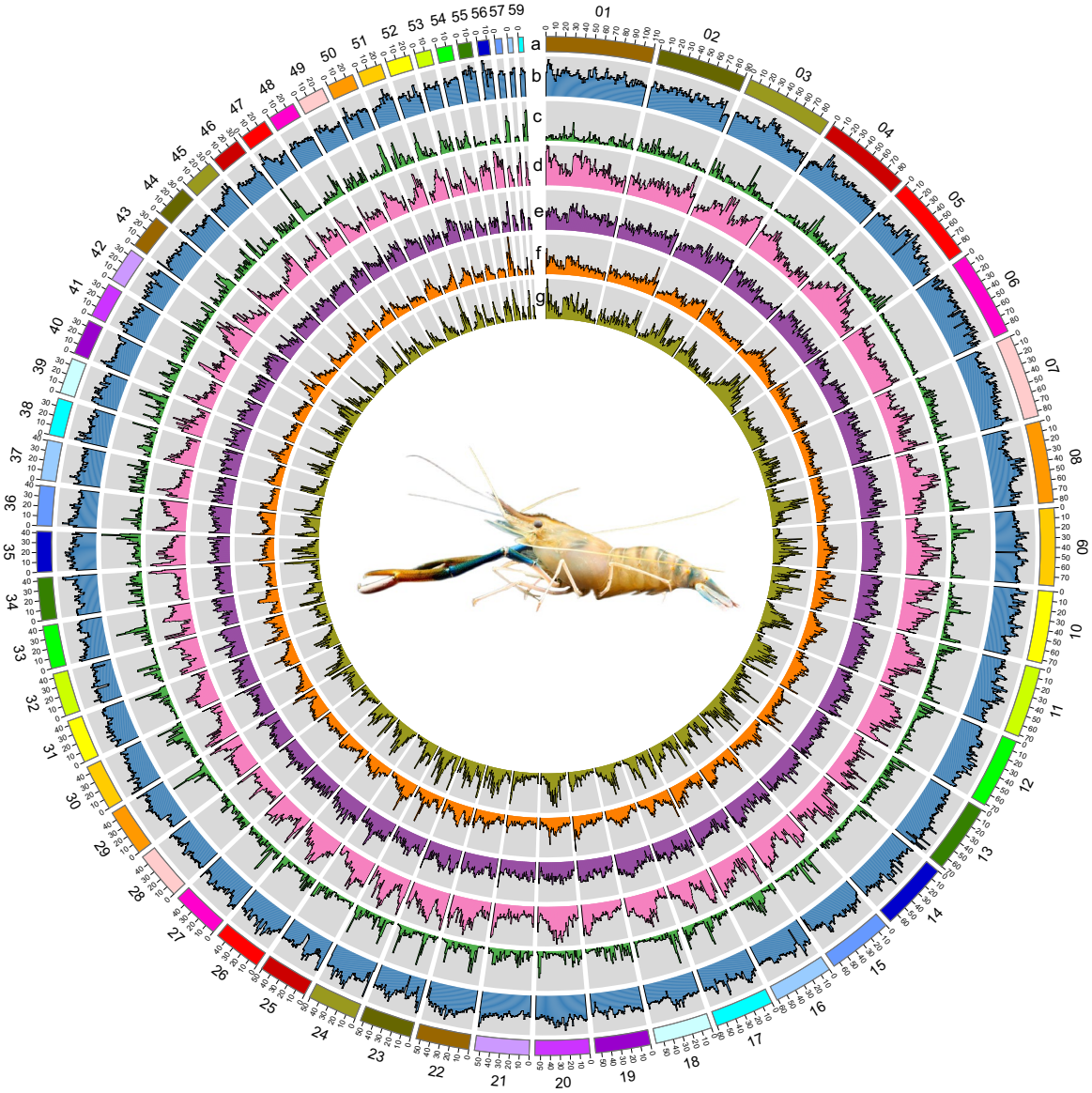


**Fig. 3** Genomic features of *Macrobrachium rosenbergii*. (**a**) The 59 pseudochromosomes; (**b**) Histogram displaying the GC content; (**c**) Gene density; (**d**–**g**) Density of the total repeat sequences (REs), DNA-REs, LINE-REs, and LTR-REs.

integrated to obtain structural information on the predicted genes. The prediction model was developed using AUGUSTUS[43], and gene structures were predicted from scratch. These prediction results were integrated using Evidence Modeler (EVM)[44] based on certain weight values (the default criterion for weight value setting is: PASA prediction, score 7 ≥ GeMoMa prediction, score 7 ≥ ab initio prediction, score 1, to obtain the initial genome gene set. Genes containing transposable elements and coding errors were removed from the initial gene set via TransposonPSI[45] comparison to obtain the final gene set.

In total, 27,111 protein-coding genes were predicted by integrating transcriptomes, homologous proteins, and de novo predictions. Each gene had an average of 5.96 exons, a mean gene length of 36580.88 bp, and a coding sequence size of 1524.77 bp. Average exon and intron lengths were 255.64 and 7061.39 bp, respectively (Table 5). The annotated protein score was 94.57% by BUSCO evaluation.

**Gene function annotation.** Gene function information, motifs, and protein domains were assigned using comparisons with public databases including SwissProt, Kyoto Encyclopedia of Genes and Genomes (KEGG)[46], eukaryotic orthologous groups (KOG)[47], gene ontology (GO)[48,49], and NCBI non-redundant protein (Nr) databases. Overall, a total of 25,470 genes, accounting for 93.95% of the predicted protein-coding genes, were successfully annotated. Specifically, the SwissProt, KEGG, KOG, GO, and Nr databases annotated approximately 62.16%, 34.75%, 49.68%, 45.83%, and 90.73% of the genes, respectively (Table 6).

| Class | Order | Super family | Number of sequences | Length of sequence (bp) | Percentage of sequence (%) |
|---|---|---|---|---|---|
| Class I | | | 3,953,846 | 841,306,550 | 28.38 |
| | LTR | | 1,799,227 | 429,324,023 | 14.48 |
| | | Copia | 62,157 | 7,008,793 | 0.24 |
| | | Gypsy | 441,001 | 128,749,094 | 4.34 |
| | | ERV1 | 251,988 | 8,381,610 | 0.28 |
| | | Unknown | 676,002 | 158,558,254 | 5.35 |
| | | Gypsy-Cigr | 12,769 | 10,962,833 | 0.37 |
| | | ERVK | 125,544 | 4,222,147 | 0.14 |
| | | DIRS | 155,822 | 96,969,060 | 3.27 |
| | | Pao | 40,844 | 13,431,145 | 0.45 |
| | | Other | 33,100 | 1,041,087 | 0.04 |
| | LINE | | 1,979,168 | 391,610,194 | 13.21 |
| | | CR1 | 43,504 | 11,507,726 | 0.39 |
| | | Penelope | 332,884 | 66,914,124 | 2.26 |
| | | L2 | 388,677 | 56,912,762 | 1.92 |
| | | Unknown | 515,255 | 121,437,607 | 4.1 |
| | | I | 66,684 | 29,130,631 | 0.98 |
| | | RTE-BovB | 168,320 | 42,379,432 | 1.43 |
| | | L1-Tx1 | 99,988 | 6,017,034 | 0.2 |
| | | L1 | 86,836 | 3,166,022 | 0.11 |
| | | CR1-Zenon | 80,073 | 22,500,566 | 0.76 |
| | | R2 | 46,118 | 3,083,969 | 0.1 |
| | | Dong-R4 | 20,540 | 6,017,253 | 0.2 |
| | | Jockey | 6,405 | 3,361,387 | 0.11 |
| | | I-Nimb | 13,726 | 9,870,786 | 0.33 |
| | | Proto2 | 8,582 | 3,543,704 | 0.12 |
| | | Other | 101,576 | 5,767,191 | 0.19 |
| | SINE | | 175,451 | 20,372,333 | 0.69 |
| | | tRNA-Core-RTE | 17,658 | 5,765,645 | 0.19 |
| | | Unknown | 90,636 | 7,499,926 | 0.25 |
| | | tRNA-L2 | 29,474 | 4,237,380 | 0.14 |
| | | Other | 37,683 | 2,869,382 | 0.1 |
| Class II | | | 5,769,144 | 531,963,681 | 17.95 |
| | DNA | | 5,299,833 | 470,119,145 | 15.86 |
| | | Unknown | 893,070 | 127,200,159 | 4.29 |
| | | Maverick | 248,163 | 12,927,591 | 0.44 |
| | | CMC-EnSpm | 821,713 | 51,791,182 | 1.75 |
| | | hAT-Charlie | 153,447 | 9,124,524 | 0.31 |
| | | MULE-MuDR | 399,053 | 74,044,674 | 2.5 |
| | | hAT-Ac | 240,846 | 10,833,507 | 0.37 |
| | | hAT-Tip100 | 165,638 | 8,436,333 | 0.28 |
| | | CMC-Transib | 144,225 | 10,453,054 | 0.35 |
| | | TcMar-Fot1 | 169,056 | 15,942,668 | 0.54 |
| | | Zisupton | 249,320 | 22,796,867 | 0.77 |
| | | P | 230,117 | 15,802,425 | 0.53 |
| | | CMC-Chapaev | 147,837 | 4,948,900 | 0.17 |
| | | Kolobok-Hydra | 60,578 | 2,973,115 | 0.1 |
| | | PIF-Harbinger | 174,807 | 7,367,268 | 0.25 |
| | | Ginger | 208,298 | 13,904,359 | 0.47 |
| | | PiggyBac | 20,419 | 6,148,723 | 0.21 |
| | | TcMar-Tc1 | 82,735 | 12,001,470 | 0.4 |
| | | Academ-1 | 28,747 | 11,791,549 | 0.4 |
| | | TcMar-Tigger | 49,572 | 11,113,135 | 0.37 |
| | | Academ | 10,524 | 5,382,439 | 0.18 |
| | | Other | 801,668 | 35,135,203 | 1.19 |
| | MITE | | 154,098 | 20,518,590 | 0.69 |

Continued

| Class | Order | Super family | Number of sequences | Length of sequence (bp) | Percentage of sequence (%) |
|---|---|---|---|---|---|
| | | Unknown | 154,098 | 20,518,590 | 0.69 |
| | RC | | 315,213 | 41,325,946 | 1.39 |
| | | Helitron | 315,213 | 41,325,946 | 1.39 |
| Total TEs | | | 9,722,990 | 1,373,270,231 | 46.33 |
| Tandem Repeats | | | 579,923 | 71,819,595 | 2.42 |
| | tandem_repeat | | 316,348 | 68,468,461 | 2.31 |
| | SSR | | 263,575 | 3,351,134 | 0.11 |
| Other | | | 167,788 | 14,383,225 | 0.49 |
| Unknown | | | 629,392 | 103,140,260 | 3.48 |
| Simple repeats | | | 51,779 | 9,765,364 | 0.33 |
| Low complexity | | | 19,618 | 3,327,843 | 0.11 |
| Total Repeats | | | 11,171,490 | 1,575,706,518 | 53.16 |

**Table 4.** Repetitive sequences and their proportions in the *M. rosenbergii* genome.

| Species | Total number of protein-coding genes | Average transcript length(bp) | Average CDS length(bp) | Average exon length(bp) | Average intron length(bp) | Genebank |
|---|---|---|---|---|---|---|
| *Macrobrachium rosenbergii* (This study) | 27,111 | 36,580.88 | 1,524.77 | 255.64 | 7,061.39 | GCA_046866055.1 |
| *Macrobrachium rosenbergii* | 22,842 | 44,842.79 | 1,574.94 | 238.19 | 7,709.90 | GCA_040412425.1 |
| *Eriocheir sinensis* | 19,445 | 20,872.81 | 1,670.51 | 228.65 | 3,045.14 | GCA_024679095.1 |
| *Macrobrachium nipponense* | 24,301 | 50,129.28 | 1,512.44 | 238.8 | 9,115.47 | GCA_015104395.2 |
| *Penaeus vanmamei* | 23,907 | 9,976.47 | 1,292.01 | 213.81 | 1,722.18 | GCA_003789085.1 |
| *Portunus trituberculatus* | 17,138 | 26,405.62 | 1,662.70 | 217.84 | 3,730.52 | GCA_017591435.1 |
| *Procambarus clarkii* | 25,734 | 26,589.27 | 1,532.83 | 266.72 | 5,278.34 | GCA_020424385.2 |
| *Scylla paramamosain* | 19,737 | 19,018.47 | 1,609.53 | 221.41 | 2,776.75 | GCA_035594125.1 |

**Table 5.** Comparison of genes identified in *M. rosenbergii* with those of other crustacean species.

| | Gene number | Percent (%) |
|---|---|---|
| Total | 27,111 | 100% |
| Swissprot | 16,851 | 62.16 |
| KEGG | 9,420 | 34.75 |
| KOG | 13,468 | 49.68 |
| GO | 12,426 | 45.83 |
| NR | 24,599 | 90.73 |
| Annotated | 25,470 | 93.95 |
| Unannotated | 1,641 | 6.05 |

**Table 6.** Functional annotation of *M. rosenbergii* genes.

## Data Records

The raw sequencing reads of all libraries were stored in the NCBI Sequence Read Archive (SRA) database (BioProject ID PRJNA1268154[50]). The assembled genome has been deposited at Genbank under the accession number GCA_046866055.1[51]. The genome annotation was deposited in the Figshare database[52].

## Technical Validation

The completeness and accuracy of the genomic assembly of *M. rosenbergii* was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.4.3)[34] with the arthropoda_odb10 database. The BUSCO assessment showed that 957 (94.37%) BUSCOs were identified as complete, of which 951 (93.88%) and 6 (0.59%) were single-copy and duplicated, respectively (Table 3), indicating a good quality of the genome. Chromosome numbers of the *M. rosenbergii* genome were confirmed by the Hi-C (Fig. 2). A protein-level BUSCO assessment using the arthropoda_odb10 library revealed 94.57% (958) completeness, with 92.79% (940) single-copy, 1.78% (18) duplicated, and 0.80% (9) fragmented orthologs, confirming high-quality gene annotation.

## Code availability

The bioinformatics analyses in this study were performed according to the manuals and protocols provided by the software developers. The corresponding parameters used by the software are described in the methods section.

Received: 7 October 2024; Accepted: 1 July 2025;
Published online: 10 July 2025

## References


1. Balazs, G. H. & Ross, E. Effect of protein source and level on growth and performance of the captive freshwater prawn, *Macrobrachium rosenbergii*. *Aquaculture.* **7**, 299–313 (1976).
2. Dong, X. *et al*. A novel virus of *Flaviviridae* associated with sexual precocity in *Macrobrachium rosenbergii*. *mSystems*. **6** (2021).
3. Niza, N. N. M. *et al*. Prevalence, molecular characterization and the effects of boyprid isopod Probopyrus buitendijki infestation on the *Macrobrachium rosenbergii* from Nyatoh River, Terengganu, Malaysia. *Regional Studies in Marine Science.* **69**, 103311 (2024).
4. Ying, N. *et al*. Transcriptome analysis of *Macrobrachium rosenbergii*: Identification of precocious puberty and slow-growing information. *Journal of Invertebrate Pathology.* **190**, 107752 (2022).
5. Phuc, T. H. *et al*. Assessment of a long-term selective breeding program for giant freshwater prawn *Macrobrachium rosenbergii* since 2007. *Aquaculture.* **541**, 736745 (2021).
6. Martínez, P. *et al*. A genome-wide association study, supported by a new chromosome-level genome assembly, suggests sox2 as a main driver of the undifferentiatiated ZZ/ZW sex determination of turbot (*Scophthalmus maximus*). *Genomics.* **113**, 1705–1718 (2021).
7. Zhang, X. *et al*. A new strain of yellow catfish carrying genome edited myostatin alleles exhibits double muscling phenotype with hyperplasia. *Aquaculture.* **523**, 735187 (2020).
8. Johnston, I. A. *et al*. Advancing fish breeding in aquaculture through genome functional annotation. *Aquaculture.* **583**, 740589 (2024).
9. Liu, H. *et al*. A genome-wide association study to identify growth-related SNPs and genes in blotched snakehead (Channa maculata). *Aquaculture Reports.* **35**, 101932 (2024).
10. Sävilammi, T. *et al*. The chromosome-level genome assembly of European grayling reveals aspects of a unique genome evolution process within salmonids. *G3 Genes|Genomes|Genetics.* **9**, 1283–1294 (2019).
11. Peng, M. *et al*. A high-quality genome assembly of the Pacific white shrimp (*Litopenaeus vannamei*) provides insights into its evolution and adaptation. *Aquaculture Reports.* **33**, 101859 (2023).
12. Huerlimann, R. *et al*. Genome assembly of the Australian black tiger shrimp (*Penaeus monodon*) reveals a fragmented IHHNV EVE sequence. *G3 Genes|Genomes|Genetics.* **12**, jkac034 (2021).
13. Xu, Z. *et al*. A chromosome-level reference genome of red swamp crayfish Procambarus clarkii provides insights into the gene families regarding growth or development in crustaceans. *Genomics.* **113**, 3274–3284 (2021).
14. Jin, S. *et al*. A chromosome-level genome assembly of the oriental river prawn, *Macrobrachium nipponense*. *GigaScience.* **10**, giaa160 (2021).
15. Chen, H. *et al*. The chromosome-level genome of *Cherax quadricarinatus*. *Scientific Data.* **10**, 313 (2023).
16. Shao, C. *et al*. The enormous repetitive *Antarctic krill* genome reveals environmental adaptations and population insights. *Cell.* **186**, 1279–1294.e1219 (2023).
17. Tang, B. *et al*. Chromosome-level genome assembly reveals the unique genome evolution of the swimming crab (*Portunus trituberculatus*). *GigaScience.* **9**, giz161 (2020).
18. Wang, J. *et al*. "Omics" data unveil early molecular response underlying limb regeneration in the Chinese mitten crab, *Eriocheir sinensis*. *Science Advances.* **8**, eabl4642 (2022).
19. Jiang, Q. *et al*. De novo transcriptome analysis of eyestalk reveals ovarian maturation related genes in Macrobrachium rosenbergii. *Aquaculture.* **505**, 280–288 (2019).
20. Peng, X. *et al*. Histopathological observation and comparative transcriptome analysis reveal immune response mechanisms to Aeromonas dhakensis infection in *Macrobrachium rosenbergii*. *Fish & Shellfish Immunology.* **142**, 109151 (2023).
21. Wang, Y. *et al*. Transcriptome analysis reveals the molecular response to salinity challenge in larvae of the giant freshwater prawn *Macrobrachium rosenbergii*. *Frontiers in Physiology.* **13**, 885035 (2022).
22. Zheng, Y. *et al*. A Chromosome-level genome assembly of giant river prawn (*Macrobrachium rosenbergii*). *Sci Data* **11**, 935 (2024).
23. Chin, C. S. *et al*. Phased diploid genome assembly with single molecule real-time sequencing. *Nature Methods.* **13**, 1050–1054 (2016).
24. Chen, S. *et al*. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* **34**, i884–i890 (2018).
25. Brown, J. *et al*. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics.* **33**, 3137–3139 (2017).
26. SMRT Link, https://www.pacb.com/support/software-downloads/.
27. Deorowicz, S. *et al*. KMC 2: fast and resource-frugal *k*-mer counting. *Bioinformatics.* **31**, 1569–1576 (2015).
28. Sun, H. *et al*. FindGSE: estimating genome size variation within human and *Arabidopsis* using *k*-mer frequencies. *Bioinformatics.* **34**, 550–557 (2017).
29. Ding, W. *et al*. PyComplexHeatmap: A Python package to visualize multimodal genomics data. *iMeta.* **2**, e115 (2023).
30. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods.* **18**, 170–175 (2021).
31. Guan, D. F. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics.* **36**, 2896–2898 (2020).
32. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods.* **9**, 357–359 (2012).
33. Burton, J. N. *et al*. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology.* **31**, 1119–1125 (2013).
34. Simão, F. A. *et al*. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* **31**, 3210–3212 (2015).
35. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics.* **32**, 2103–2110 (2016).
36. Wang, X. & Wang, L. GMATA: An integrated software package for genome-scale ssr mining, marker development and viewing. *Frontiers in Plant Science.* **7**, 1350 (2016).
37. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research.* **27**, 573–580 (1999).
38. Abrusán, G. *et al*. TEclass-a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics.* **25**, 1329–1330 (2009).
39. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics.* **4**, 10 (2009).
40. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* **29**, 15–21 (2012).
41. Haas, B. J. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research.* **31**, 5654–5666 (2003).
42. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research.* **44**, e89–e89 (2016).
43. Stanke, M. *et al*. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics.* **24**, 637–644 (2008).

44. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidence Modeler and the program to assemble spliced alignments. *Genome Biology*. **9**, R7 (2008).
45. Urasaki, N. *et al*. Draft genome sequence of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and subtropical regions. *DNA Research*. **24**, 51–58 (2017).
46. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. **28**, 27–30 (2000).
47. Galperin, M. Y. *et al*. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*. **43**, D261–D269 (2015).
48. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nature Genetics*. **25**, 25–29 (2000).
49. Zdobnov, E. M. & Apweiler, R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. **17**, 847–848 (2001).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP588275 (2025).
51. Wang, Y. Macrobrachium rosenbergii isolate YW-2024, whole genome shotgun sequencingproject. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046866055 (2025).
52. Wang, Y. A high-quality chromosome-level genome assembly and annotation of the giant freshwater prawn (*Macrobrachium rosenbergii*). *Figshare* https://doi.org/10.6084/m9.figshare.25928944 (2024).


## Acknowledgements

This research was funded by the Scientific Innovation Fund, PRFRI (2023CXYC1), the special fund for seed industry revitalization project of Guangdong Province (2024-SPY-00-007, 2023-SJS-00-001), the Central Public-Interest Scientific Institution Basal Research Fund, CAFS (2022SJ-XT1, 2023TD38), the National Freshwater Genetic Resource Centre (FGRC18537), the China-ASEAN Maritime Cooperation Fund (CAMC-2018F) and the Guangdong Rural Revitalization Strategy Special Provincial Organization and Implementation Project Funds (2022-SBH-00-001).

## Author contributions

Y.W. and X.Z. designed the study. L.Y., H.L., J.W., K.H., H.L., Q.Z., Q.S., F.L., W.Z. and C.L. performed the experiments and analyzed the data. L.Y. and Y.W. wrote the paper. Y.W. and X.Z. revised the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to X.Z. or Y.W.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.